DOCUMENT RESUME

ED 334 982 IR 015 145

AUTHOR Frick, Theodore W.
TITLE A Comparison of an Expert Systems Approach to Computerized Adaptive Testing and an Item Response Theory Model.
PUB DATE 91
NOTE 22p.; In: Proceedings of Selected Research Presentations at the Annual Convention of the Association for Educational Communications and Technology; see IR 015 132.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Comparative Analysis; *Computer Assisted Testing; *Expert Systems; *Item Response Theory; Mastery Tests; *Models; Test Construction; Test Items; Test Length; Test Validity

ABSTRACT

Expert systems can be used to aid decisionmaking. A computerized adaptive test is one kind of expert system, although not commonly recognized as such. A new approach, termed EXSPRT, was devised that combines expert systems reasoning and sequential probability ratio test stopping rules. Two versions of EXSPRT were developed, one with random selection of items (EXSPRT-R) and one with intelligent selection (EXSPRT-I). Two empirical studies were conducted in which these two new methods were compared to the traditional SPRT and to an adaptive mastery testing (AMT) approach based on item response theory (IRT). The EXSPRT-I tended to be more efficient than the AMT, EXSPRT-R, and SPRT models in terms of average test lengths. Although further research is needed, the EXSPRT-I initially appears to be a strong alternative to both IRT- and SPRT-based adaptive testing when categorical decisions about examinees are desired. The EXSPRT-I is clearly less complex than IRT, both conceptually and mathematically. It also appears to require many fewer examinees to establish empirically a rule base when compared to the large numbers needed to eliminate parameters for item response functions in the IRT model. (20 references) (Author/BBM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED384982

Title:

A Comparison of an Expert Systems Approach to Computerized Adaptive Testing and an Item Response Theory Model

Author:

Theodore W. Frick

120145

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Michael R. Simonson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ABSTRACT

Expert systems can be used to aid decision making. A computerized adaptive test is one kind of expert system, though not commonly recognized as such. A new approach, termed EXSPRT, was devised that combines expert systems reasoning and sequential probability ratio test stopping rules. Two versions of EXSPRT were developed, one with random selection of items (EXSPRT-R) and one with intelligent selection (EXSPRT-I). Two empirical studies were conducted in which these two new methods were compared to the traditional SPRT and to an adaptive mastery testing (AMT) approach based on item response theory (IRT). The EXSPRT-I tended to be more efficient than the AMT, EXSPRT-R and SPRT models in terms of average test lengths. Although further research is needed, the EXSPRT-I initially appears to be a strong alternative to both IRT- and SPRT-based adaptive testing when categorical decisions about examinees are desired. The EXSPRT-I is clearly less complex than IRT, both conceptually and mathematically. It also appears to require many fewer examinees to establish empirically a rule base when compared to the large numbers required to estimate parameters for item response functions in the IRT model[1].

THEORETICAL ISSUES

An Overview of Expert Systems

One of the more practical results from extant research in artificial intelligence is the application of expert systems reasoning to aid in decision making or problem solving. Expert systems have been developed, for example, to help physicians identify types of bacterial infections, to aid investor decisions on buying and selling stock, for aid in assembling components of computer systems, for making decisions about where to drill for oil, for assisting underwriters in making insurance policies, and for diagnosing causes of equipment failures to help repairpersons (cf., [1]).

An expert system consists of a set of production rules or frames, often called a 'knowledge base'. The name, 'expert system', was coined because a knowledge base is typically constructed by interviewing one or more experts in some domain of knowledge. An attempt is made to capture their reasoning processes, when they solve problems in that knowledge domain, in the form of "If..., then..." rules. For example, in MYCIN, a famous early expert system for diagnosing bacterial infections, one of the rules is:

IF 1) the gram stain of the organism is negative, and
2) the morphology of the organism is rod, and
3) the aerobicity of the organism is anaerobic,

---

[1]This study was supported in part by a grant from the Proffitt Foundation, School of Education, Indiana University. Hing-Kwan Luk adapted and extended considerably the author's computer code for conducting the test re-enactments with the four methodologies. Luk also assisted with statistical analyses. Thomas Plew must be acknowledged for his significant contribution of an initial method of intelligent item selection for the EXSPRT, subsequently refined by the author. Had Tom not asked the questions he did in a doctoral seminar taught by the author in 1987, then EXSPRT might not exist today.

THEN there is suggestive evidence (.7) that the identity of the organism is Bacteroides. ([2], p. 34)

This particular rule is one of over 400 such rules that comprise the MYCIN knowledge base. A computer program, called an 'inference engine', uses this rule set as data to help physicians identify unknown bacteria. The program makes inferences by using both the rule set and specific answers to questions it asks the physician about properties of the current situation (e.g., patient symptoms, white blood cell count, and other lab test results). MYCIN has been shown to be more accurate in its identifications of bacteria than typical practicing physicians, particularly in identifying those bacteria which are rarely observed.

Expert systems are not usually viewed as replacements for human decision makers, but as aids or tools for such persons. Expert systems obviously cannot perform in areas not covered by the knowledge base. Furthermore, decisions reached by expert systems can be no better than the accuracy of the knowledge or rules that comprise the database.

In education and training, expert systems principles have been applied mostly in intelligent tutoring systems ([3], [4]). As an example, GUIDON was later developed from MYCIN in an attempt to teach physicians how to identify different kinds of bacteria [5].

Similarities between Expert Systems and Adaptive Tests

One efficient and empirically validated approach to computerized adaptive testing (CAT) is based on item response theory (e.g., [6]). An adaptive test is no longer than necessary to obtain a satisfactory estimate of an examinee's ability, and items are selected which are close to his or her estimated ability level. For example, if a person misses a question, a somewhat easier question is next asked. On the other hand, if a question is answered correctly, then a slightly more difficult question is subsequently selected. A computerized adaptive test does not waste time administering questions that are too hard or too easy for a particular individual. Adaptive tests tend to be shorter than conventional fixed-length tests and the results are as reliable if not more so (e.g., [6]).

Expert systems and adaptive computer-based tests have many properties in common:

1. The rule base for a CAT is a set of item characteristic curves estimated from prior test administrations. That is, each item characteristic curve is a compact way of saying, "If the examinee ability level is *X*, and item *Y* is asked, then the probability of a correct response is predicted to be *Z*."

2. Both expert systems and CATs use inference engines that are often Bayesian or Bayesian-like. Even if rules do not have probabilities (or confidence factors) associated with them, they can still be treated as a special Bayesian case where associated probabilities are either one or zero (cf., [7]).

3. The goal of an expert system is to choose from a number of alternatives (e.g., causes of equipment failure) using the rule base and answers to questions it selects and asks of a particular user. The goal of a CAT is to estimate an examinee's achievement or ability level with enough precision to make a decision such as pass/fail or a grade classification using a rule base of item characteristic curves and answers to questions it selects and gives to examinees.

4. An expert system selects which questions it asks by using forward or backward chaining and the rule base. A CAT can select questions on the basis of the amount of information they provide, depending on the ability of an examinee. For example, Weiss and Kingsbury use a maximum information search and selection (MISS) procedure [6].

Thus, although not widely recognized at this time, an adaptive testing system is one type of an expert system. The author realized this when developing computer code for an expert system, having already developed code for Bayesian decision methodologies and a computer-based testing system.

On hindsight, expert systems and adaptive tests have much in common. Yet in the research literature it appears that these two threads of development have been almost entirely independent. One camp has grown out of an artificial intelligence movement and the other from a psychological testing and measurement perspective. A recent computer search of numerous bibliographic databases only turned up thirteen articles where the terms, 'expert systems' or 'artificial intelligence' and 'adaptive' or 'computer' and 'testing' or 'test' were used as descriptors. The two camps not only use different language to describe their activities, but also tend to publish in different journals and attend different conferences.

THE DEVELOPMENT OF *EXSPRT*

A problem with the IRT-based approach to adaptive testing faced by many practitioners, however, is that a relatively large number of examinees must be tested in advance in order to estimate accurately item parameters of difficulty, discrimination, and lower asymptotes (200 to 1000 depending on the model used and the number of items in a pool). Furthermore, proponents of the Rasch model (one-parameter IRT model) have indicated that there is no valid way of estimating item discrimination and lower asymptotes for the two- and three-parameter models without imposing arbitrary constraints (cf., [8]).

The author has previously investigated the predictive validity of the sequential probability ratio test (SPRT) for making mastery decisions, where the lengths of tests were adapted according to student performance ([9]). Mastery decisions reached with the SPRT, when used conservatively, agreed highly with those based on total test results. Nonetheless, the SPRT does not explicitly take into account variability in item difficulty, discrimination or chances of guessing as does the three-parameter IRT model. Moreover, items are selected randomly in the SPRT, rather than on the basis of their characteristics and estimated examinee ability or achievement level as in the MISS procedure.

Is there some middle ground between the relatively simplistic SPRT decision model and the relatively sophisticated IRT-based approach? When considering the problem from an expert systems perspective, a solution became apparent. Instead of considering a continuum of alternatives, as is the case in IRT-based CAT, it was hypothesized that if the goal of an adaptive testing system is to choose between a few discrete alternatives (e.g., mastery or nonmastery; grades of A, B, C, etc.), then it should be possible to develop a satisfactory rule base from a smaller sample of examinee test data--compared to the IRT model.

Development of the Rule Base for a Given Test Item Pool

Assume that we have developed a pool of test items which match a particular instructional objective and that our goal is to choose between two alternatives for any given student: mastery or nonmastery of the objective. For each item *i* in the pool we create four rules:

Rule i.1: If the examinee is a *master* and item *i* is selected, then the probability of a correct response is $P(C_i \mid M)$.

Rule i.2: If the examine is a *master* and item *i* is selected, then the probability of an incorrect response is $P(\neg C_i \mid M)$.

Rule i.3: If the examinee is a *nonmaster* and item *i* is selected, then the probability of a correct response is $P(C_i \mid N)$.

Rule i.4: If the examine is a *nonmaster* and item *i* is selected, then the probability of an incorrect response is $P(\neg C_i \mid N)$.

Notice that these rules are essentially in the same if-then form as the sample rule from the MYCIN expert system at the beginning of this article. In MYCIN the rules were developed on the basis expert knowledge on the co-occurrence of various kinds of gram stain, morphology and aerobicity and the incidence of each kind of bacterium. In our case, we will rely on empirical data collected on the test items with a representative sample of examinees who are characterized by the discrete categories from among which our expert system will later attempt to choose.

In the dichotomous case, the estimates of probabilities of correct responses to items by masters and nonmasters are determined as follows[2]:

1. Give the pool of test items to a representative group of examinees, about half of whom are expected to be masters and half nonmasters--i.e., for whom you expect a wide range of scores on the test.

2. Choose a mastery cut-off score (e.g., .85).

3. Divide the original group into a mastery group and nonmastery group based on their total test scores and the mastery cut-off.

4. For each item in the mastery group, estimate the probabilities of correct and incorrect responses by the following formulas (see [10])[3]:

$$P(C_i \mid M) = (\#r_{im} + 1)/(\#r_{im} + \#w_{im} + 2) \qquad \{1.1\}$$

$$P(\neg C_i \mid M) = 1 - P(C_i \mid M) \qquad \{1.2\}$$

---

[2]Note that this reasoning can be extended to more than two categories, such as letter grade designations.

[3]Note that the estimates of these probabilities of correct responses to items by masters and nonmasters will never be one or zero. This means that, in the EXSPRT Bayesian updating process during the administration of a test to an examinee, the probabilities of the mastery and nonmastery alternatives will never be zero or one, though these extremes may be closely approached.

where $\#r_{im}$ = number of persons in the mastery group who answered the item correctly;

and $\#w_{im}$ = number of persons in the mastery group who missed the item.

5. Do likewise for the nonmastery group for each item:

$$P(C_i \mid N) = (\#r_{in} + 1)/(\#r_{in} + \#w_{in} + 2) \quad \{1.3\}$$

$$P(\neg C_i \mid N) = 1 - P(C_i \mid N) \quad \{1.4\}$$

Method of Making Inferences in the EXSPRT

The reasoning procedure employed in our expert systems approach is Bayesian, with the addition of stopping rules from the sequential probability ratio test (SPRT) (cf., [9, 10, 11]). After each observation (i.e., administration of a test item), a likelihood ratio is computed:

$$LR = \frac{P_{om} \prod_{i=1}^{n} P(C_i|M)^s \, [\, 1-P(C_i|M) \,]^f}{P_{on} \prod_{i=1}^{n} P(C_i|N)^s \, [\, 1-P(C_i|N) \,]^f} \quad \{2\}$$

where

P_{om} = prior probability that the examinee is a master,

P_{on} = prior probability that the examinee is a nonmaster[4]

and $s = 1, f = 0$ if item i is answered correctly,

or $s = 0, f = 1$ if item i is answered incorrectly.

The three stopping rules are:

If $LR \geq (1 - \beta) \div \alpha$, then stop asking questions and choose mastery. {3.1}

If $LR \leq \beta \div (1 - \alpha)$, then stop asking questions and choose nonmastery. {3.2}

Else ask another question, update LR, and reiterate rules 3.1 to 3.3. {3.3}

---

[4]Note that if the prior probabilities of mastery and non-mastery are equal, then they drop out of the formula for the likelihood ratio.

Alpha and beta are Type I and II decision errors, respectively. Alpha is the probability of choosing mastery when the nonmastery alternative is actually true. Beta is the probability of choosing nonmastery when the mastery alternative is true.

For numerical examples of this Bayesian reasoning process, the reader is referred to [12]. Also note that expert systems can use reasoning procedures other than Bayesian, such as the Dempster-Shafer rules. See [7] for further examples of Bayesian reasoning in expert systems.

Random Item Selection: EXSPRT-R

When the EXSPRT was initially conceived, it was seen as an extension of the Bayesian approach to the SPRT, as noted in [9], using empirically derived data for estimating the probabilities of correct responses by masters and nonmasters to each test item rather than *average* probabilities across all items. In this initial approach to the EXSPRT, items were selected randomly without replacement, and it was assumed that observations were independent in order to multiply the conditional probabilities to form the likelihood ratio. This version of the EXSPRT is referred to as EXSPRT-R, in contrast an intelligent item selection procedure discussed below.

A reader who is knowledgeable about expert systems may note that the method of choosing questions in the EXSPRT approach to adaptive testing *departs* from the typical method of forward or backward chaining used in many expert systems. One might argue that the EXSPRT approach is not really expert systems' reasoning because forward or backward chaining is apparently not used to determine which question to ask next in order to reach, or move closer to, a conclusion (i.e., choose from a number of discrete alternative goals).

On the other hand, one who understands expert systems deeply, having written computer code to carry out the reasoning, will conclude that the EXSPRT is doing essentially the same thing as occurs in either backward or forward chaining. The rules in the EXSPRT are all of the same form, however, in contrast to rule bases in expert systems where a variable's value in a consequent condition in one rule is also part of an antecedent condition in another rule, etc. (cf. [7]). One could say that instead of just asking a particular question once, and then immediately making an inference based on the answer to that question–as is usually the case in most expert systems–in the EXSPRT many questions of the same kind are asked until an inference can be made confidently. In the EXSPRT-R questions are chosen at random in order to comprise a representative sample from the item pool.

Intelligent Item Selection: EXSPRT-I

Thomas Plew was not satisfied with EXSPRT-R, since it did not use information about test items in the selection process [13]. This stimulated the joint development with the present author of an item selection procedure that is modeled after basic principles used by Weiss and Kingsbury in the MISS (maximum information search and selection) procedure. Though the principles are comparable, the mathematical approaches are quite different.

In the EXSPRT-I (i.e., with "intelligent" item selection), the reasoning is as follows:

Item discrimination. If we are trying to choose between mastery or nonmastery alternatives, then an item is more discriminating when the difference between probabilities of correct responses by masters and nonmasters is greater. For example,

if the probabilities of a correct response to item #5 are .90 for masters and .25 for nonmasters, then item #5 is very discriminating (difference = .65). On the other hand, if the probability of a correct response to item #53 is .85 for masters and .75 for nonmasters, then this item is much less discriminating (difference = .10). Or if the probability of a correct response to item #12 is .60 for masters and .80 for nonmasters, then such an item is negatively discriminating (difference = -.20).

Thus, the discrimination index for item *i* is defined:

$$D_i = P(C_i \mid M) - P(C_i \mid N) \qquad \{4\}$$

Item/examinee incompatibility. Not only do we want to select highly discriminating items, but also we want to select items that are matched to an examinee's estimated achievement or ability level. In theory, we gain little additional information by administering items which are very easy or very hard for a given individual. Better items would be those which a person has a 50/50 chance of answering correctly--i.e., which are very close to her or his achievement level. For example, if an examinee's achievement level is estimated to be .80 (on a scale from zero to one), then a good item would be one that was answered *incorrectly* by 80 percent of the examinees in the item parameter estimation sample [$P(C_i) = .20$ for masters and nonmasters combined].

Thus, the item/examinee incompatibility index is defined for each item:

$$I_{ij} = abs\{(1 - P(C_i)) - E(\Phi_j)\} \qquad \{5\}$$

where $$E(\Phi_j) = (\#r_j + 1)/(\#r_j + \#w_j + 2) \qquad \{6\}$$

and $$P(C_i) = (\#r_i + 1)/(\#r_i + \#w_i + 2) \qquad \{7\}$$

Note that $\#r_j$ and $\#w_j$ are the numbers of questions answered correctly and incorrectly, respectively, thus far in the test by the current examinee. Note also that the estimate of $P(C_i)$ is based on the *total* number of persons in the parameter estimation sample for item *i*, irrespective of mastery status. Thus, $\#r_i$ is the number of persons who answered item *i* correctly and $\#w_i$ is the number who answered it incorrectly. Finally, note that the item/examinee incompatibility index is based on the absolute value of the difference between the estimate of the probability of an *incorrect* response to the item and the estimate of the current examinee's achievement level (proportion correct metric).

Item utility. As a test proceeds, item utilities are re-calculated for all items remaining in the pool, in order to select and administer a new one that now has the most utility for an examinee:

$$U_{ij} = D_i/(I_{ij} + \delta) \qquad \{8\}$$

where δ = some arbitrary small constant (e.g., .0000001), to prevent division by zero in case $I_{ij} = 0$.[5]

Thus, each utility value is simply the ratio of the discrimination of item *i* and its incompatibility with person *j*'s achievement level. The item that is selected next in the EXSPRT-I (intelligent selection) is the remaining one with the greatest utility at that point for that particular examinee. This means that the item selected next is the one which *discriminates best* between masters and nonmasters *and* which is *least incompatible* with the current estimate of that examinee's achievement level. Note that item utilities change during a test, depending on an examinee's performance which affects the estimate of his/her achievement level in the item/examinee incompatibility index. In effect, the EXSPRT-I is comparable to the two-parameter item response theory model (IRT--see [14]) in that both item discrimination and item difficulty are considered in the item selection process.

Unanswered Questions

Since the EXSPRT-R and EXSPRT-I are new approaches to computerized adaptive testing, two empirical studies were conducted to compare these approaches to extant IRT-based adaptive mastery testing and SPRT approaches (cf., [6], [9], [14]). Of major concern was the accuracy with which each adaptive model could predict decisions based on total test scores. Does each adaptive method make mastery and nonmastery decisions with no more errors than would be expected by *a priori* error rates? Second, how efficient is each adaptive method in terms of average test lengths for mastery and nonmastery decisions? Are any of the methods more efficient than others?

FIRST STUDY

Digital Authoring Language Test

A computer-based test on the structure and syntax of the Digital Authoring Language was constructed, consisting of 97 items, and referred to as the DAL test. This test was comprised of multiple-choice, binary-choice, and short-answer questions. The test was highly reliable (Cronbach α = .98). The DAL test was also very long, usually taking between 60 and 90 minutes to complete, and it was very difficult for most examinees (mean score = 63.2 percent correct, S.D. = 24.6).

Examinees

The persons who took the DAL test were mostly either current or former graduate students in a course on computer-assisted instruction taught by the author. Those students who were currently enrolled at the time took the DAL test twice, once about mid-way through the course when they had some knowledge of DAL--which they were required to learn for developing CAI programs--and once near the end of the course when they were expected to be fairly proficient in DAL. The

[5]Alternatively, δ_i could be considered as some kind of "guessing" factor for the item. However, this will not be considered in the present paper.

remainder of the examinees took the DAL test once. Since the test was long and difficult, no one was asked to take the test who did not have some knowledge of DAL or other authoring languages.

Test Administration

The DAL test was individually administered by the Indiana Testing System [15]. As an examinee sat at a computer terminal, items were selected at random without replacement from the total item pool until all items were administered. Students were not allowed to change previous answers to questions, nor was feedback given during the test. Upon completion of test, complete data records were stored in a database, including the actual sequence in which items were randomly administered to a student, response time, literal response to each item, and the item scoring (correct or incorrect). Examinees were informed of their total test scores at the end of the test. There were a total of 53 administrations of the DAL test in the first study.

Experimental Methods

The basic procedure was to re-enact each test, using actual examinee responses in the database, for each of the four adaptive methodologies: 1) IRT-based adaptive mastery testing (AMT--with maximum information search and selection [MISS]), 2) sequential probability ratio test (SPRT), 3) EXSPRT-R (random selection of items), and 4) EXSPRT-I (intelligent selection of items--see above descriptions).

Item parameter estimation. Two random samples of examinees were used to estimate item parameters (n = 25 and n = 50), the latter containing the former. This was done to see if increasing the sample size used for parameter estimation would result in fewer decision errors in the four methods. Due to the relatively small sample sizes, the one-parameter AMT model was used--i.e., only b_i estimates were obtained for the two samples using program BICAL [16]. For the EXSPRT-R and EXSPRT-I, the rule base for each parameter estimation sample was constructed using formulas {1.1}, {1.2}, {1.3} and {1.4}. The mastery cut-off was set at 72.5 percent, half way between the established .85 mastery level and .60 nonmastery level used in an earlier study of the SPRT only [9]. In the current study, however, the mastery and nonmastery levels for the SPRT were established *empirically* from the .725 cut-off and the two parameter estimation samples. The mean proportion correct for masters was used as the mastery level and the mean proportion correct for nonmasters was used as the nonmastery level in each sample. In effect, the SPRT was treated just like the EXSPRT-R, except that the rule quadruplets for all items were the same in the SPRT, based on the sample means for masters and nonmasters, respectively.

Test re-enactments. Once the parameter estimation samples were chosen, then two doctoral assistants independently wrote computer programs in two different languages (Pascal and DAL) to construct the rule bases for the EXSPRT, and to carry out the four different adaptive testing methods on the same 53 sets of test administrations. This was done to reduce the possibility of error in coding these rather complex methodologies, especially the AMT model. When results did not agree, as was occasionally the case, this helped to identify and ameliorate errors in coding. The one difference that was not correctable was traced to the precision of arithmetic in DAL and Pascal on a VAX minicomputer.

It was discovered that on occasion the MISS procedure in the two programs would begin to select different items in the AMT model after 15 to 20 items had been

retroactively "administered" to an examinee. This occurred because the updating of the estimate of Θ and its variance, and in turn the item information estimates for that Θ estimate, would tend to differ very slightly in the two code versions as a test progressed. Consequently, the MISS procedure would occasionally pick a different item in the two different versions when estimates of item information were very close for two or more items remaining in the pool. From that point on in a test, different item sequences were observed. The average AMT test length in the DAL version tended to be about one item shorter, compared to the Pascal version, but the decisions reached were the same with one exception.

These discrepancies do point out a problem inherent in the IRT-based approach, which contains numerous multiplications, divisions, and exponentials (see [14], formulas {9) to {25}). Very small errors due to rounding or differences in precision of arithmetic can magnify themselves rather quickly. This problem was not observed with the EXSPRT-I, EXSPRT-R, or SPRT—other than differences in the millionth's decimal place when computing probability ratios.

1. *AMT re-enactment.* The mastery cut-off was converted to Θ_c using the test characteristic curve (see [14], formula {24}) and the item parameter database constructed from the respective parameter estimation sample (either n = 25 or 50). The value of Θ_c was used as the initial prior Θ and the prior variance was set to one, as recommended by Weiss and Kingsbury [9]. The MISS procedure was used to select the next test item for the re-enactment for each examinee (see [14], formulas {10} to {12}). The correctness of the examinee's response to that item was determined by retrieving it from the database. Bayesian updating of Θ and its variance was accomplished with Owen's method [17] (see [14], formulas {13} to {22}). After each item was "administered", the AMT stopping rules were applied using a .95 confidence interval (see [14], formulas {25.1} to {25.3}). If a decision could be reached, the re-enactment was ended at that point. The number of questions answered correctly and incorrectly in the AMT and the decision reached for that examinee were written to a computer data file. Also stored in that file were the total test score for that examinee and the agreement between the AMT decision and the total test decision. If no decision could be reached by the AMT model before exhausting the test item pool, then a decision was forced at the end of the test: If the current estimate of Θ was greater than or equal to Θ_c, the examinee was considered to be a master; otherwise a nonmaster.

2. *SPRT.* The mastery and nonmastery levels required by the SPRT were empirically established from the parameter estimation samples, as described above. Since the SPRT requires random selection of items, test items were "administered" in a random order. Alpha and β levels were set at 0.025, to make the overall decision error rate (.05) equivalent to the .95 confidence interval method used in the AMT approach. When the SPRT reached a mastery or nonmastery decision, results were stored in a separate data file in the same manner as described above for the AMT.

3. *EXSPRT-R.* As in the SPRT, items were "administered" in a random order. However, the rule bases constructed from the parameter estimation samples were used, of course, in the EXSPRT-R method of Bayesian updating (formula {2}, with equal prior probabilities) and SPRT stopping rules ({3.1} to {3.3}). For a description of EXSPRT-R procedures, see [12] for an example of expert systems reasoning during computer-based testing. When the EXSPRT-R reached a decision, the test re-enactment was ended and results written to a data file as before.

4. *EXSPRT-I.* This method was the same as the EXSPRT-R, except that items were selected intelligently, based on their utility indices (see formulas {4} to {8}). Thus, like the AMT, items were not "administered" randomly for each re-enactment. Since no feedback was given during the test it is unlikely that decisions reached by both AMT and EXSPRT-I methods would be systematically affected by factors other than differences in the adaptive methods themselves. One mitigating factor might be examinee fatigue, where they were more likely to answer questions incorrectly at the end of the long and difficult test. However, since all test items were originally administered in a different random order for each individual, it is very unlikely that fatigue would systematically bias any findings.

Results from the First Study

For the DAL test, IRT item parameters (b_i's) were estimated from samples of 25 and 50 examinees. EXSPRT rule bases were also derived from the same samples. Descriptive information is given about the two samples in the left side of Table 1. It can be seen that there were about the same proportions of masters and nonmasters in each sample. In the sample of 50 there were 23 masters whose average test score was 87.3 percent, and 27 nonmasters who scored 45.1 percent correct.

Mean test lengths of each of the four methods, variation in test lengths, and decision accuracies were compared. If the decision made by an adaptive method was the same as that reached on the basis of the entire test item pool, this was considered to be a "hit". Thus, the accuracy measures are the percent of correct predictions made by each method. There were 28 nonmasters and 25 masters identified by the entire 97-item test, when the cut-off score was set at 72.5 percent correct.

First, note that the parameter sample size seems to make little difference in the mean test length *within* each method. For example, within the AMT model 20.6 items were required for nonmastery decisions when item parameters were based on a sample of 25, compared to a mean of 18.3 for the sample of 50. For the EXSPRT-I, 5.6 items were required for nonmastery decisions in the sample of 25, compared to a mean of 5.9 for the parameter sample of 50. Please note--and this is confusing--that the mean test lengths for each of the four methods are based on the same 53 test administrations, where all 97 items were originally given, and which were re-enacted under each adaptive method. The size of the *parameter estimation* sample refers to the number of examinees randomly selected on whom the item difficulties were estimated for the AMT model and on whom the item rule bases were constructed for the EXSPRT-R and EXSPRT-I models.

Decision accuracies. For the 53 administrations of this DAL test there does seem to be some difference in decision accuracies within each model for the two parameter estimation sample sizes. The decision accuracies tended to be high for all methods. Decision accuracies were compared to expected values of .975 correct mastery decisions and .975 correct nonmastery decisions, using Chi-square goodness of fit tests (e.g., see [18]). A significant Chi-square ($p < .05$) means that the observed decision accuracies *departed* from what was expected according the *a priori* decision error rates that were established for each of the four adaptive testing methods.

When 25 examinees were used for parameter estimation, there were two significant departures from expected accuracy. The EXSPRT-I was 85.7 percent accurate in nonmastery decisions, which significantly differed from the expected 97.5

Table 1. Efficiency and Accuracy of the Four Adaptive Testing Methods in the First Study.[6]

Item Parameter Sample Description		ADAPTIVE TESTING METHOD			
		AMT	SPRT	EXSPRT-R	EXSPRT-I
	Mean Score (S.D.) n	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy
Masters	87.46 (7.90) 12	8.40 (9.62) 100.0	8.72 (5.16) 92.0	7.56 (3.22) 100.0	5.44 (1.23) 100.0
Nonmasters	42.66 (15.83) 13	20.57 (24.45) 96.4	10.54 (7.14) 85.7*	12.71 (15.46) 96.4	5.64 (2.02) 85.7*
Total	64.16 (25.99) 25	14.83 (19.77) 98.1	9.68 (6.29) 88.7	10.28 (11.65) 98.1	5.55 (1.68) 92.5
Masters	87.27 (7.89) 23	8.28 (8.19) 100.0	10.36 (6.92) 96.0	8.44 (5.74) 96.0	6.84 (2.64) 100.0
Nonmasters	45.06 (16.25) 27	18.29 (24.43) 96.4	10.11 (10.97) 89.3*	9.39 (9.15) 92.9	5.93 (2.28) 92.9
Total	64.47 (24.89) 50	13.57 (19.14) 98.1	10.23 (9.20) 92.5	8.94 (8.94) 94.3	6.36 (2.47) 96.2

*Percent accuracies were tested by goodness of fit, where .975 accuracy was expected according to the a priori error rates for masters and nonmasters. Only those percent accuracies which differed significantly from the expected accuracies, according to a chi-square test (d.f. = 1, $p < .05$) are marked with an asterisk.

[6]Alpha = ß = 0.025 for the SPRT, EXSPRT-R, and EXSPRT-I; a .95 confidence interval was used with the AMT. There were 53 administrations of the DAL test which were re-enacted for each of the four adaptive methods.

percent accuracy. At the same time, however, the EXSPRT-I was reaching decisions when the other models were requiring two to four times as many items. The SPRT accuracy for nonmastery decisions was also significantly lower than expected.

When 50 examinees were used for parameter estimation, the AMT, EXSPRT-R, and EXSPRT-I models were within the expected range of accuracy. The SPRT failed to make as many correct nonmastery decisions as were expected. What is notable is how well *all* of the adaptive methods predicted total test decisions, while using between 5 and 20 items from the 97-item pool to reach those decisions--a very substantial reduction in test lengths (95 to 80 percent decrease).

Efficiency. A repeated measures ANOVA (MANOVA) was conducted to see if there were significant differences among the mean test lengths for the four adaptive methods. This was done for the results based on the parameter sample of 50 for the 53 test administrations. Hotelling's T^2 was significant at the .05 level. However, the sphericity assumption was violated, due to the large differences in variances among the four methods. A *post hoc* comparison procedure suggested by Marascuilo and Levin ([19], pp. 373-381) for this kind of situation was conducted for all pair-wise contrasts of mean test lengths. One statistically significant difference was found. The mean test length for the SPRT was significantly greater than that for the EXSPRT-I. Even though some of the other contrasts have greater magnitudes of difference, the within-method variances are very different themselves. It can be noted that, overall, the AMT model required about twice as many items to reach decisions (13.6) as did the EXSPRT-I (6.4), though it was not statistically significant at the .05 level.

The variances in average test lengths within each adaptive method were significantly different, as noted above in violation of the sphericity assumption. The variance in test lengths for the AMT model was approximately 60 times larger than that for the EXSPRT-I model (19.14^2 vs. 2.47^2). In the AMT model, tests tended to be longer before nonmastery decisions were reached, and there was much more variation in test lengths compared to the remaining models. The variation in lengths of tests with EXSPRT-I method was relatively small compared to variation in the remaining models.

SECOND STUDY

Computer Functions Test

A computer-based test on how computers work, consisting of 85 items, was constructed. The COM test, as it is referred to here, was comprised of about half multiple-choice, one-fourth binary choice, and one-fourth fill-in type questions (Cronbach α = .94). Compared to the DAL test, the COM test was much easier for most examinees (mean score = 79.0 percent, S.D. = 13.6).

Examinees

About half of those who took the COM test were from two sections of an introductory graduate-level course on use of computers in education. The remainder were mostly volunteers from an undergraduate-level course for non-education majors who were learning to use computers. A small number of students were volunteers recruited at the main library on campus.

Test Administration and Experimental Methods

The COM test was individually administered by the Indiana Testing System in the same manner as the DAL test. There were a total of 104 administrations of the COM test in the second study. The same four adaptive testing methods were re-enacted from actual examinee test data in the very same manner as described above for the DAL test.

Results from the Second Study

Since there were more administrations of the COM test, parameter estimation samples of 25, 50, 75 and 100 were selected at random. Four sets of b_i coefficients were obtained for the AMT model and four rule bases were constructed for the EXSPRT models based on the same four parameter estimation samples. See the left sides of Tables 2.1 and 2.2 for descriptive information about the parameter estimation samples.

Accuracy of predictions. When the parameter estimation sample was 25, all four adaptive methods did not perform as well as expected in correctly predicting nonmasters in the 104 administrations of the COM test. Chi-square goodness of fit tests showed that all four methods significantly departed from the expected accuracy rates. EXSPRT-I had the worst accuracy, but it should be noted that there were only *seven* nonmasters in the estimation sample for creating the rule base, so this is not surprising.

When the parameter estimation sample was 50, the AMT and EXSPRT-I models still made significantly fewer correct nonmastery decisions than expected *a priori*. On the other hand, the SPRT and EXSPRT--both of which use random selection of items vs. intelligent selection in the AMT and EXSPRT-I--predicted masters and nonmasters correctly within the bounds of expected error rates.

When the parameter estimation sample was 75 (55 masters and 20 nonmasters when the cut-off was 72.5 percent correct), all models predicted well except the AMT, which made significantly fewer correct nonmastery decisions than were expected *a priori*.

When the parameter estimation sample was 100, the AMT model still had problems with accuracy of nonmastery classifications. And strangely enough, the AMT model also made significantly fewer correct *mastery* decisions than were expected. The SPRT, EXSPRT-R, and EXSPRT-I all correctly predicted masters and nonmasters within the bounds of expected accuracies. It should be noted that it is generally recommended that a minimum of 200 examinees be used for estimating b_i parameters in the IRT-based, one-parameter AMT model. Only half that number were available in this study. Thus, it is not surprising that the AMT model performed less well than it should, since estimation of the item difficulty parameters was not as precise as desired.

Efficiency. Average test lengths of the four adaptive methods were compared for the 100 examinee parameter estimation situation only. See the bottom half of Table 2.2. A MANOVA again revealed that the sphericity assumption was violated, and so the same procedure as described above for the DAL test was used in post hoc comparisons of the adaptive COM test length means [19].

Table 2.1. Efficiency and Accuracy of the Four Adaptive Testing Methods in the Second Study.[7]

Item Parameter Sample Description		ADAPTIVE TESTING METHOD			
		AMT	SPRT	EXSPRT-R	EXSPRT-I
	Mean Score (S.D.) n	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy
Masters	86.21 (6.35) 18	8.37 (13.59) 97.4	11.71 (7.80) 98.7	10.05 (5.42) 98.7	4.57 (3.60) 98.7
Nonmasters	48.07 (9.10) 7	33.93 (31.83) 82.1*	14.39 (15.81) 85.7*	15.00 (10.41) 82.1*	7.07 (2.36) 67.9*
Total	75.53 (18.84) 25	15.25 (23.02) 93.3	12.43 (10.55) 95.2	11.38 (7.39) 94.2	5.24 (3.49) 90.4
Masters	87.16 (5.68) 35	11.83 (18.23) 94.7	15.08 (9.06) 96.1	11.71 (8.28) 98.7	5.72 (3.92) 96.1
Nonmasters	53.65 (10.44) 15	31.89 (29.97) 78.6*	17.39 (14.50) 92.9	15.82 (12.70) 96.4	7.93 (6.21) 89.3*
Total	77.11 (17.15) 50	17.23 (23.61) 90.4	15.70 (10.77) 95.2	12.82 (9.78) 98.1	6.32 (4.72) 94.2

*Percent accuracies were tested by goodness of fit, where .975 accuracy was expected according to the a priori error rates for masters and nonmasters. Only those percent accuracies which differed significantly from the expected accuracies, according to a chi-square test (d.f. = 1, $p < .05$) are marked with an asterisk.

[7]Alpha = ß = 0.025 for the SPRT, EXSPRT-R, and EXSPRT-I; a .95 confidence interval was used with the AMT. There were 104 administrations of the COM test which were re-enacted for each of the four adaptive methods.

Table 2.2. Efficiency and Accuracy of the Four Adaptive Testing Methods in the Second Study (cont'd).

Item Parameter Sample Description		ADAPTIVE TESTING METHOD			
		AMT	SPRT	EXSPRT-R	EXSPRT-I
	Mean Score (S.D.) n	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy
Masters	87.68 (5.93) 55	10.21 (16.96) 94.7	16.64 (10.35) 97.4	11.78 (6.34) 97.4	7.70 (7.13) 94.7
Nonmasters	56.00 (10.17) 20	28.93 (28.58) 82.1*	16.29 (16.96) 92.9	14.75 (15.54) 100.0	7.82 (4.85) 100.0
Total	79.23 (15.84) 75	15.25 (22.21) 91.3	16.55 (12.39) 96.2	12.58 (9.71) 98.1	7.73 (6.57) 96.2
Masters	87.47 (6.33) 75	13.75 (20.80) 93.4*	16.97 (10.75) 96.1	13.58 (9.51) 98.7	7.64 (6.12) 94.7
Nonmasters	56.00 (11.34) 25	31.50 (29.77) 78.6*	13.04 (10.66) 92.9	12.32 (10.78) 96.4	8.93 (7.41) 100.0
Total	79.60 (15.77) 100	18.53 (24.70) 89.4	15.91 (10.82) 95.2	13.24 (9.83) 98.1	7.99 (6.48) 96.2

*Percent accuracies were tested by goodness of fit, where .975 accuracy was expected according to the a priori error rates for masters and nonmasters. Only those percent accuracies which differed significantly from the expected accuracies, according to a chi-square test (d.f. = 1, $p < .05$) are marked with an asterisk.

When nonmastery decisions were made, the AMT model required significantly longer tests than either the SPRT, EXSPRT-R or EXSPRT-I. The AMT model required about 32 items to reach nonmastery decisions, compared to the EXSPRT-I, which required about nine items. Moreover, the AMT made significantly fewer correct nonmastery decisions than expected, as noted above. When mastery decisions were reached, test lengths for the SPRT and EXSPRT-R methods (15 and 12) were significantly longer than the EXSPRT-I (6 items). Mean test lengths for mastery decisions in the AMT and EXSPRT-I models were not significantly different at the .05 level.

When looking at decisions overall, the following contrasts were significantly different: the AMT, SPRT, and EXSPRT-R methods each required significantly longer tests than did the EXSPRT-I model. The AMT model required over twice as many items as did the EXSPRT-I (19 vs. 8).

Summary. It would appear from the COM test data that the EXSPRT-I is significantly more efficient than the other adaptive methods. Indeed, it is rather remarkable that the EXSPRT-I can make such highly accurate mastery and nonmastery decisions with relatively few test questions. It is also notable that the EXSPRT-R and SPRT also made highly accurate predictions, but were less efficient than the EXSPRT-I. The AMT performed worst of all, not only resulting in longer adaptive tests but also in making significantly more prediction errors than theoretically expected.

DISCUSSION

Adaptive tests tended to be shorter with the DAL test than with the COM test. See Tables 1, 2.1 and 2.2. Of the 53 administrations of the DAL test, there were 28 nonmasters and 25 masters when the cut-off was set at 72.5 percent and when examinees answered all 97 items. The overall average test score was 63.2 (S.D. = 24.6). In the second study with the COM test there were 104 administrations of this test, with 76 masters and 28 nonmasters when the entire 85-item test was taken (grand mean = 79.0, S.D. = 13.6, mastery cut-off = 72.5 percent).

A similar study was conducted by Plew [13] with yet a different test on computer literacy (referred to as the LIT test here). In his sample of 183 examinees there were 54 masters and 129 nonmasters based on total test results from the 55-item pool. The cut-off for this test was 59.5 percent, and the overall average score was 51.5 percent (S.D. = 14.2).

One thing that appears to affect the average test lengths is the location and shape of the distribution of examinee achievement levels in relation to the cut-off selected. In the first study, the distribution was somewhat bimodal and relatively flat, with about half the examinees scoring above and below the cut-off. In the second study, the distribution was positively skewed, with about three-fourths of the examinees scoring above the 72.5 percent cut-off on the entire test item pool. In the Plew study, over two-thirds of the examinees were classified as nonmasters on the entire 55-item test [13]. The distribution of this group was close to normal, with the mean being about 8 percentage points below the selected cut-off.

The author has previously conducted a number of computer simulations comparing the three-parameter AMT model with the SPRT and a third adaptive method based on Bayesian posterior beta distributions ([14],[20]). One important

finding in those studies was that none of the adaptive methods performed as well as expected–and average test lengths tended to be longer–when the distribution of examinees was mostly clustered around the cut-off. Adaptive tests were shorter and accuracies agreed with theoretical expectations when the distributions of examinee achievement levels were much flatter. The same phenomenon appears to have occurred in the present two empirical studies, as well as in Plew's.

The second factor that may affect results is the number of test items in each pool and their properties. When there are more test items, and there are more items available at each ability or achievement level, then both the AMT and EXSPRT-I tend to be more efficient and more accurate. In both adaptive methods which rely on "intelligent" selection of items, Bayesian posterior estimates are affected more dramatically when there are highly discriminating items available whose difficulty levels are close to the current estimate of an examinee's achievement level. A real problem occurs with smaller item pools, as was the case with the LIT test in Plew's study: After the best items have been administered early in a test, the remaining items tend to provide little additional information. That is, there are diminishing returns after some point because there are no really appropriate items left.

SUMMARY

Expert systems can be used to aid decision makers. A computerized adaptive test (CAT) is one kind of expert system, though not commonly recognized as such. When item response theory is used in a CAT, then the knowledge or rule base is a set of item characteristic curves (ICC's).

Normally an expert system consists of a set of questions and a rule base. An inference engine uses answers to the questions and the rule base to choose from a set of discrete alternatives. If an adaptive test is viewed this way, then it is possible to construct "If ..., then ..." rules about test items that are not functions, as are ICC's. A new approach, termed EXSPRT, was devised that combines expert systems reasoning and sequential probability ratio test stopping rules. EXSPRT-R uses random selection of test items, whereas EXSPRT-I incorporates an intelligent selection procedure based on item utility coefficients.

These two new methods were compared to the traditional SPRT and to an IRT-based approach to adaptive mastery testing (AMT). Two empirical studies with different tests and types of examinees were carried out.

In the first study the EXSPRT-I model required about half as many items as did the AMT approach (6 vs. 14), though the difference was not statistically significant. When 50 examinees were used for item parameter estimation and rule base construction, all four methods (AMT, SPRT, EXSPRT-R and EXSPRT-I) made highly accurate mastery and nonmastery decisions.

In the second study the EXSPRT-I method again required about half as many items as did the AMT model (8 vs. 19), and this time the difference was statistically significant. When 100 examinees were used for estimation purposes, the SPRT, EXSPRT-R, and EXSPRT-I correctly predicted masters and nonmasters within the bounds of the expected theoretical error rates. The AMT model, however, made significantly more prediction errors than expected.

Although further research is needed, the EXSPRT-I initially appears to be a strong alternative to both IRT- and SPRT-based adaptive testing when categorical

decisions about examinees are desired. The EXSPRT-I is clearly less complex than IRT, both conceptually and mathematically. It also appears to require many fewer examinees to establish empirically a rule base when compared to the large numbers required to estimate parameters for item response functions in the IRT model.

On the other hand, the EXSPRT is vulnerable, as is classical test theory, in that a representative sample of examinees must be selected for constructing rule quadruplets. This seems to be a small price to pay for the advantages of theoretical parsimony and operational efficiency.

REFERENCES

1. Winston, P. and Prendergast, K. (1984). *The AI business: The commercial uses of artificial intelligence*. Cambridge, MA: The MIT Press, 17-40.
2. Davis, R. (1984). Amplifying expertise with expert systems. In P. Winston and K. Prendergast (Eds.), *The AI business: The commercial uses of artificial intelligence*. Cambridge, MA: The MIT Press, 17-40.
3. Kearsley, G. (Ed.), (1987). *Artificial intelligence and instruction: Applications and methods*. Reading, MA: Addison-Wesley.
4. Sleeman D. and Brown, J. (Eds.) (1982). *Intelligent tutoring systems*. New York, NY: Academic Press.
5. Clancey, W. (1987). Methodology for building an intelligent tutoring system. In G. Kearsley (Ed.), *Artificial intelligence and instruction: Applications and methods*. Reading, MA: Addison-Wesley, 193-227.
6. Weiss, D. and Kingsbury, G. (1984). Application of computerized adaptive testing to education problems. *Journal of Educational Measurement, 21*, 361-375.
7. Heines, J. (1983). Basic concepts in knowledge-based systems. *Machine-Mediated Learning, 1*(1), 65-95.
8. Wright, B. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14(2), 97-116.
9. Frick, T. (1989). Bayesian adaptation in computer-based tests and computer-guided practice exercises. *Journal of Educational Computing Research*, 5(*1*), 89-114.
10. Schmitt, S. (1969). *Measuring uncertainty*. Reading, MA: Addison-Wesley.
11. Wald, A. (1947). *Sequential analysis*. New York: Wiley.
12. Frick, T. (1990). Analysis of patterns in time: A method of recording and quantifying temporal relations in education. *American Educational Research Journal*, *27*(1), 180-204.
13. Plew, G. T. (1989). *A comparison of major adaptive testing strategies and an expert systems approach*. Bloomington, IN: Doctoral dissertation, Indiana University Graduate School.
14. Frick, T. (1990). A comparison of three decision models for adapting the length of computer-based mastery tests. *Journal of Educational Computing Research*, *6*(4), 479-513.

15. Frick, T. (1986). The Indiana Testing System (ITS, Version 1.0). Bloomington: Department of Instructional Systems Technology, School of Education, Indiana University.
16. Mead, R., Wright, B., and Bell, S. (1979). BICAL (Version 3). Chicago: Department of Education, University of Chicago.
17. Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.
18. Glass, G. and Hopkins, K. (1984). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice-Hall.
19. Marascuilo L. and Levin, J. (1983). *Multivariate statistics in the social sciences: A researcher's guide.* Monterey, CA: Brooks/Cole.
20. Frick, T., Luk, H.-K., and Tyan, N.-C. (1987). *A comparison of three adaptive decision-making methodologies used in computer-based instruction and testing.* Bloomington, IN: Final Report, Proffitt Foundation, Indiana University School of Education.